Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072126
Vol. 7, No. 3, Juni 2020, hlm. 501-510 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

KOMBINASI *K-MEANS* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK MEMPREDIKSI UNSUR SARA PADA *TWEET*

Wiga Maulana Baihaqi*[1], Muliasari Pinilih[2], Miftakhul Rohmah[3]

[1]Teknologi Informasi, Universitas Amikom Purwokerto
[2,3]Sistem Informasi, Universitas Amikom Purwokerto
Email: [1]wiga@amikompurwokerto.ac.id, [2]mpinilih@amikompurwokerto.ac.id, [3]m.miftakhul.mi@gmail.com
*Penulis Korespondensi

(Naskah masuk: 25 Juni 2019, diterima untuk diterbitkan: 22 April 2020)

Abstrak

Tulisan yang disampaikan melalui twitter dinamakan dengan *tweets* atau dalam bahasa indonesia lebih dikenal dengan kicau, tulisan yang di*share* memiliki batas maksimum, tulisan tidak boleh lebih dari 140 karakter, karakter disini terdiri dari huruf, angka, dan simbol. Penyalahgunaan dalam berpendapat sering terjadi di media sosial, sering kali pengguna media sosial dengan sadar atau tidak sadar telah membuat konten yang mengandung isu Suku (dalam hal ini menyangkut keturunan), agama, ras (kebangsaan) dan antargolongan (SARA). Perlu adanya analisis yang dapat mengidentifikasi secara otomatis apakah kalimat yang ditulis pada media sosial mengandung unsur SARA atau tidak, akan tetapi korpus tentang kalimat yang mengandung unsur SARA belum ada, selain itu label kalimat yang menandakan kalimat SARA atau bukan tidak ada. Penelitian ini bertujuan untuk membuat *corpus* kalimat yang mengandung unsur SARA yang didapatkan dari twitter, kemudian melabeli kalimat dengan label mengandung unsur SARA dan tidak, serta melakukan *sentiment* klasifikasi. Algoritme yang digunakan untuk proses pelabelan adalah k-*means*, sedangkan *Support Vector Machine* (SVM) digunakan untuk proses klasifikasi. Hasil yang diperoleh berdasarkan k-*means* antara lain 118 *tweet* positif SARA dan 83 *tweet* negatif SARA. Dalam proses klasifikasi menggunakan dua metode validasi, yaitu 5-*fold cross validation* yang dibandingkan dengan 10-*fold cross validation*, hasil akurasi dari kedua metode validasi tersebut yaitu, masing-masing 64,18% dan 63,68%. Berdasarkan hasil akurasi yang diperoleh untuk meningkatkan hasil akurasi, data hasil proses k-*means* diolah kembali dengan validasi pakar bahasa, hasil yang diperoleh menjadi 139 *tweet* positif SARA dan 62 *tweet* negatif SARA, hasil akurasi meningkat menjadi 70,15% dan 71,14%. Dari hasil yang didapatkan, twitter dapat dijadikan sumber untuk membuat *corpus* mengenai kalimat SARA, dan metode yang diusulkan berhasil untuk proses pelabelan dan sentimen klasifikasi, akan tetapi masih perlu peningkatan hasil akurasi.

Kata kunci: *twitter, k-means, support vector machine, SARA*

K-MEANS AND SUPPORT VECTOR MACHINE (SVM) COMBINATION TO PREDICT SARA ELEMENTS ON TWEET

Abstract

Posts sent via twitter are called tweets or in Indonesian better known as chirping, the posts shared have a maximum limit, the writing cannot be more than 140 characters, the characters here consist of letters, numbers, and symbols. Broadcasting in discussions that often occur on social media, often users of social media consciously or unconsciously have created content that contains issues of ethnicity, religion, race (nationality) and intergroup (SARA). Obtained from the analysis that can automatically contain sentences on social media containing no SARA or not, but the corpus about sentences containing SARA does not yet exist, other than that the sentence label indicates SARA or no sentence. This study aims to make sentence corpus containing SARA elements obtained from twitter, then label sentences with labels containing elements of SARA and not, and conduct group sentiments. The algorithm used for the labeling process is k-means, while Support Vector Machine (SVM) is used for the classification process. The results obtained based on k-means include 118 positive SARA tweets and 83 negative SARA tweets. In the classification process using two validation methods, namely cross-fold validation of 5 times compared with 10-fold cross validation, the accuracy of the two validation methods is 64.18% and 63.68%, respectively. Based on the results obtained to improve the results, the k-means process data were reprocessed with linguists, the results obtained were 139 positive SARA tweets and 62 SARA negative tweets, the results of which increased to 70.15% and 71.14%. From the results obtained,

Twitter can be used as a source to create a corpus about SARA sentences, and methods that have succeeded in labeling and classification sentiments, but still need to improve the results of accuracy.

Keywords: *twitter, k-means, support vector machine, SARA*

1. PENDAHULUAN

Media sosial dalam perkembangan teknologi informasi memiliki andil besar bagi manusia dalam memberikan kemudahan untuk bersosialisasi. Media sosial merupakan media yang dapat diakses secara langsung melalui *internet*, artinya masyarakat yang menggunakan media tersebut dapat melakukan berbagai aktivitas seperti *sharing* (berbagi), *participating* (berpartisipasi) dan *creating* (membuat) konten-konten seperti wiki, forum, blog, dan jejaring sosial lainnya. Tentu saja semua ini tak terlepas dari kemajuan teknonologi melalui aplikasi-aplikasi yang tersedia di *internet* (Tim Pusat Humas Kementerian Perdagangan RI, 2014).

Menurut (Hootsuite dan We Are Social, 2018) dilihat dari lama pengaksesannya, pengguna *internet* di Indonesia memenfaatkan waktu untuk berselancar di dunia maya selama 8,85 jam, hal tersebut belum termasuk untuk berselancar di jejaring sosial yang biasanya memakan waktu selama 3,38 jam dalam 24 jam. Sedangkan dilihat dari jumlah pengguna, perbandingan penguna *internet* dengan seluruh jumlah penduduk di Indonesia adalah 1:2, artinya setengah dari jumlah penduduk di Indonesia merupakan pengguna internet, 130 juta diantaranya pengguna aktif media sosial dengan penetrasi 49 persen. Merebaknya situs media sosial yang muncul banyak mendorong pada hal-hal baru sehingga bermanfaat bagi kehidupan manusia menjadi lebih mudah, efektif dan efisien. Penggunaan media sosial di Indonesia seperti pada Gambar 1.

Sumber : (Hootsuite dan We Are Social, 2018)

Gambar 1. Penggunaan Media Sosial di Indonesia

Gambar 1 menunjukkan aplikasi media sosial yang digunakan oleh masyarakat Indonesia (Hootsuite dan We Are Social, 2018), Youtube menempati posisi pertama dengan persentase 43 %, posisi ke dua Facebook dengan persentase 41%, posisi ke tiga Whatsapp dengan persentase 40%, posisi ke empat Instagram dengan persentase 38%, posisi ke lima Line dengan persentase 33%, posisi ke enam BBM dengan persentase 28%, posisi ke tujuh Twitter dengan persentase 27%, dan lain-lain.

Tulisan yang disampaikan melalui twitter dinamakan dengan *tweets* atau dalam bahasa indonesia lebih dikenal dengan kicau, tulisan yang di*share* memiliki batas maksimum, tulisan tidak boleh lebih dari 140 karakter, karakter disini terdiri dari huruf, angka, dan simbol (Tim Pusat Humas Kementerian Perdagangan RI, 2014). Menurut data di situs Semiocast 30 Juli 2012, semiocast telah menganalisis 517 juta profil pengguna twitter yang dibuat sebelum 1 Juli 2012. Indonesia menempati posisi ke lima dunia dengan jumlah *tweet* sebanyak 29,4 juta orang. Jakarta menempatkan kota paling aktif jumlah *tweet* yang di-*posting* dengan teknologi semiocast, 27% dari semua *tweet* geo-lokasi disemua kota. Lebih dari 2 % dari semua *tweet* di*posting* di ibukota Indonesia. Aktifnya penggunaan twitter di Indonesia pengguna memiliki kebebasan berpendapat dalam menyampaikan opini, membagikan informasi yang kini sering digunakan untuk mengkritik masyarakat hingga menyampaikan kritik kepada pimpinan daerah.

Suku (dalam hal ini menyangkut keturunan), agama, ras (kebangsaan) dan antargolongan (SARA) merupakan masalah yang sering terjadi di masyarakat. SARA terjadi akibat adanya kebebasan dalam berpendapat namun kebebasan ini disalahgunakan sehingga menyinggung pihak-pihak tertentu yang termasuk dalam SARA tersebut. Akibat yang terjadi adalah munculnya kriminalitas seperti tindak kekerasan, diskriminasi dan pelecehan (Rudybyo, 2011). Penyalahgunaan dalam berpendapat sering terjadi di media sosial, sering kali pengguna media sosial dengan sadar atau tidak sadar telah membuat konten yang mengandung isu Suku (dalam hal ini menyangkut keturunan), agama, ras (kebangsaan) dan antargolongan (SARA). Sehingga Kementrian Komunikasi dan Informatika (Kementerian Komunikasi dan Informatika, 2017) pada tahun 2017 dapat merekap kasus konten di media sosial yang mengandung unsur ketidaksukaan terhadap pihak atau golongan tertentu sebanyak 13.829, selain itu terdapat 6.973 yang mengandung konten tidak sesuai dengan fakta yang ada.

Analisis sentimen atau *opinion mining* merupakan suatu bidang studi yang mengacu secara luas berdasarkan komputasi linguistik, *text mining* dan pengolahan bahasa alami yang bertujuan untuk menganalisa pendapat, sikap, emosi, penilaian, sentimen serta evaluasi seseorang yang didasarkan

apakah pembicara atau penulis berkenan dengan suatu layanan, organisasi, individu, tokoh publik, acara ataupun kegiatan lain (Liu, 2012).

Pada umumnya, sentimen analisis terdiri dari tiga langkah utama: *feature extraction*, *sentiment clustering*, dan *sentiment classification*. *Feature extraction* merupakan langkah untuk memperoleh fitur pada data berupa teks yang tidak terstruktur. Kemudian *sentiment clustering* dilakukan untuk mengelompokan teks yang sejenis, ini dilakukan apabila data yang digunakan belum memiliki label atau *class*. Begitu juga dengan data yang akan digunakan dalam penelitian ini, data yang digunakan pada penelitian ini adalah kalimat yang mengandung unsur SARA, data tersebut tidak memiliki label bahkan belum terdapat korpus.

Pelabelan k-*means* telah menjadi salah satu algoritme pengelompokan yang paling mudah beradaptasi selama bertahun-tahun. Keunggulan dengan efisiensi tinggi dan kemampuan menangani dimensi tinggi menjadikan algoritme k-*means* pilihan yang tepat untuk analisis sentimen. Biasanya, rentang nilai antara 3 dan 7 dipilih untuk jumlah *cluster* (Han, Kamber & Pei, 2012; Muzakir, 2014). Penelitian (Xu dan Tian, 2015; Mustakim, 2012) menunjukkan k-*means* memiliki kompleksitas waktu yang relatif rendah (O (nkd)) yang menunjukkan kinerja yang lebih cepat. Ini juga bekerja dengan baik dengan kumpulan data besar. Padahal itu cukup sensitif terhadap data yang tidak bersih. Dengan pertimbangan profesional, para peneliti menyimpulkan bahwa ini akan menjadi pendekatan yang baik. Penelitian terbaru (Korovkinas, Danėnas & Garšva, 2019; Garay, Yap & Sabellano, 2019; Liu dan Lee, 2018) juga menggunakan k-*means* untuk proses *labeling* teks sebelum proses *sentiment classifications*.

Setelah mendapatkan label dari masing-masing *text*, langkah selanjutnya biasanya proses klasifikasi. *Support Vector Machine* atau lebih sering disebut dengan SVM merupakan algoritme yang sering kali digunakan oleh peneliti untuk proses mengklasifikasikan data berupa *text*, SVM dapat dengan baik membedakan kelas atau label yang dimiliki oleh masing-masing text (Prasetyo, 2014). Sebuah studi empiris tentang perbandingan antara lima pendekatan pembelajaran yang diawasi membuktikan SVM menjadi pilihan yang menguntungkan untuk kerangka kerja yang diusulkan. SVM adalah penggolong linier dengan mengubah representasi fitur yang dinormalisasi menjadi vektor, koefisien linier dengan dimensi yang sama dalam ruang fitur menggunakan prediktor linier (*hyperplane*) (Medhat, Hassan & Korashy, 2014). Inti dari SVM adalah untuk mentransfer data dimensi tinggi ke vektor dimensi rendah. Di bidang analisis sentimen, SVM adalah salah satu pendekatan pembelajaran mesin yang paling mudah beradaptasi.

Terbukti bahwa dalam penelitian (Rahmawati, Marjuni & Zeniarja, 2017), k-*means* dan SVM digunakan untuk menganalisis pendapat dari pengguna media sosial twitter mengenai terselenggaranya pemilihan kepala daerah yang dilaksanakan secara bersama-sama, hasil yang diperoleh menunjukkan bahwa 82% algoritme yang digunakan dapat memprediksi pendapat pengguna sosial twitter dengan tepat. Penggunaan algoritme k-*means* dan SVM juga dibuktikan oleh penelitian (Somantri, Wiyono dan Dairoh, 2016), kombinasi metode tersebut terbukti lebih akurat dibandingkan hanya menggunakan algoritme SVM dalam membantu mahasiswa dalam memilih tema skripsi.

Dari permasalahan yang ada dan beberapa penelitian terdahulu menganai analisis sentimen dan klasifikasi data berupa teks, maka penelitian ini akan mengumpulkan korpus mengenai kalimat dari twitter yang mengandung unsur SARA. Dan pada penelitian ini mengusulkan metode k-*means* untuk proses *labeling* dan SVM dalam proses klasifikasi apakah konten pada twitter memiliki unsur SARA atau tidak.

2. METODE PENELITIAN

2.1. Dataset

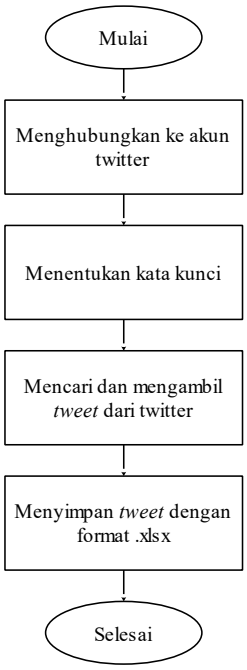
Dalam penelitian ini data yang digunakan adalah *tweet* berbahasa Indonesia yang menyinggung unsur SARA. Contoh data *tweet* yang digunakan sebagai berikut:

Tabel 1. Data *Tweet*

Tweet
Isu sara atau blunder mulut gubernur sebelumnya? Di Solo non muslim bisa jadi walikota di Kalbar non muslim juga jadi gubernur. Sebagai warga DKI Jakarta saya jengah dengan tuduhan isu sara.
Target liberal untuk membolehkan orang kafir memimpin di wilayah mayoritas umat muslim #propagandaliberal
2-mk mohon ahok tidak arogan dalam memerintah. Kasihan dengan cina2 lainnya yang miskin, baik dan tidak salah jika mereka jadi korban.
Oleh karena itu, Quran bukan kitab suci bukan pula menyebabkan kita tahu untuk menggaulinya. Nabi Muhammad bukan pula manusi suci

2.2. Pengumpulan Data

Dalam penelitian ini data sekunder diambil berdasarkan *tweet* berbahasa Indonesia yang menyinggung unsur SARA. Data dikumpulkan dengan cara *crawling* ke twitter, *tool* yang digunakan untuk proses data *crawling* adalah RapidMiner, dan operator yang digunakan adalah *search twitter*, *search* twitter digunakan sebagai penghubung ke akun twitter. Konektor twitter memungkinkan untuk dengan mudah mengakses data Twitter langsung dari RapidMiner Studio. Konektor dapat mencari frasa, *tweet*, atau informasi profil pengguna. Pada penelitian ini data yang dibutuhkan yaitu hanya tweet. Gambar 2 menunjukkan langkah *crawling* data *tweet*.



Gambar 2. Langkah *crawling* data *tweet*

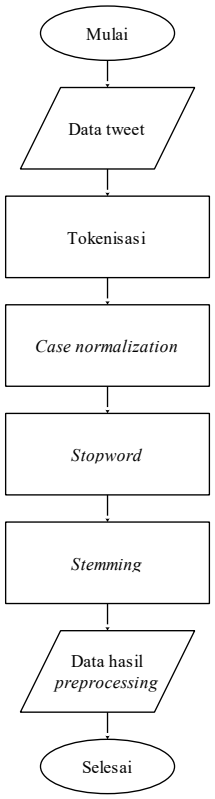
Berdasarkan Gambar 2, hal pertama yang harus dilakukan adalah menghubungkan ke akun twitter yang aktif untuk mendapatkan hak akses. *Connector* Twitter menggunakan mekanisme otentikasi yang disebut OAuth 2.0. Alih-alih memberikan RapidMiner nama pengguna dan kata sandi, maka akan menghasilkan token akses yang dapat digunakan oleh RapidMiner Studio untuk terhubung ke akun Twitter pengguna. Token ini tidak dapat digunakan oleh aplikasi lain dan membantu menjaga kredensial Twitter pengguna tetap aman.

Kemudian menentukan kata kunci, dalam penelitian ini menentukan kata kunci yang berkaitan dengan agama, ras, dan suku. Secara otomatis *tool* yang digunakan akan mencari data *tweet* berdasarkan kata kunci yang telah ditentukan. Terakhir yaitu menyimpan *tweet* yang talah didapatkan. Data *tweet* yang dikumpulkan disimpan dalam bentuk *file* bertipe .xlsx, agar data yang akan digunakan ke dalam RapidMiner pada *read excel* dapat terbaca dengan baik.

2.3. *Preprocessing* Data

Dalam proses ini data yang telah didapat dilakukan proses *preprocesing* untuk memilih data yang dapat digunakan dan menghilangkan data yang tidak diperlukan untuk tahap analisis. Tahapan yang dilakukan pada proses *preprocesing* ditunjukkan pada Gambar 3.

Langkah pertama pada tahap *preprocessing* adalah tokenisasi, tokenisasi merupakan suatu proses untuk memecahkan teks ke dalam bentuk kata. Data *tweet* yang sudah dipecah menjadi kata-kata kemudian diubah menjadi huruf kecil semua atau *lowercase,* langkah tersebut disebut dengan *case normalization* (Rozi, Arianto & Hapsari, 2019). Kemudian kata yang sering muncul tetapi tidak memiliki pengaruh apapun dalam analisis sentimen harus dihilangkan, proses ini dinamakan *stopword* (Somantri dan Apriliani, 2018). Langkah terakhir pada *preprocessing* data adalah *stemming*, merupakan proses mengurangi imbuhan pada kata untuk bentuk kata dasar, jadi hasil akhir pada tahap *preprocessing* adalah *tweet* dalam bentuk kata-kata dasar.



Gambar 3. Langkah *preprocessing* data

2.4. Pembobotan Kata

Proses pembobotan adalah proses merubah *text* ke bentuk numerik. Pada penelitian ini menggunakan *term frequency-inverse document frequency* (tf-idf). Algoritme TF-IDF adalah metode untuk membuat vektor kalimat berdasarkan frekuensi kata tersebut muncul (Ma dan Chen, 2019). Representasi dokumen tf-idf adalah skema pembobotan istilah umum dalam pengambilan informasi, yang juga telah ditemukan efektif untuk klasifikasi dokumen (Zhang, Yoshida & Tang, 2011; Manevitz dan Yousef, 2001; Ke *et al*., 2011). Ini mewakili dokumen menggunakan vektor dengan dimensi sebagai ukuran kosa kata dari korpus dan elemen yang sesuai dengan bobot tf-idf dari setiap kata w dalam dokumen d.

2.5. K-*Means Clustering*

Proses *clustering* untuk memberikan kelas pada sebuah kalimat pada *tweet*. Dalam proses ini menentukan kelas tiap kalimat menggunakan bantuan kamus dengan algoritma *k-means* untuk menentukan golongan kelas positif mengandung

SARA atau negatif mengandung SARA. Langkah-langkah yang dilakukan pada *k-means* adalah inisialisasi dan iterasi (Duwairi dan Abu-Rahmeh, 2015; Gulnashin, Sharma & Sharma, 2019). Proses inisialisasi dilakukan untuk menempatkan semua data *tweet* pada satu kelompok secara acak kemudian menghitung *centroid* sedangkan proses iterasi menghitung kembali jarak terpendek setiap data *tweet* secara berulang-ulang hingga *centorid* yang dihasilkan dalam iterasi saat ini tidak berubah atau sama dengan semua *centroid* yang dihasilkan pada iterasi sebelumnya.

2.6. Klasifikasi

Dalam proses klasifikasi digunakan algoritma *Support Vector Machine* (SVM). SVM bertujuan menemukan *hyperplane* (batas keputusan) terbaik yang memisahkan dua kelas yaitu positif mengandung SARA atau negatif mengandung SARA (Joachims, 1998; Gao, Cheng & Yu, 2019). Batas *hyperplane* diukur dan titik maksimalnya diukur untuk menemukan pembagi dua kelas. Jarak yang memisahkan *hyperplane* pada dua kelas itulah yang disebut dengan batas, sedangkan istilah *support vector* merupakan data yang mimiliki jarak paling minimum dengan batas.

2.7. Validasi dan Evaluasi

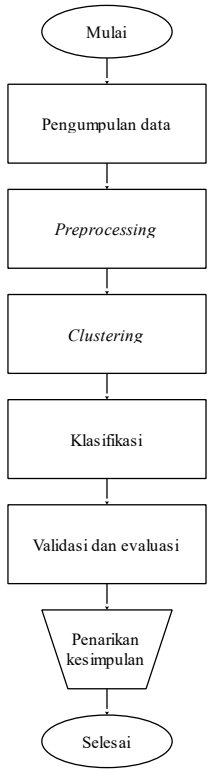
Diperlukan dua metode, baik untuk mengklasifikasikan dan memprediksi *tweet* yang berisi SARA, yaitu validasi silang 5 kali lipat dan validasi silang 10 kali lipat. Validasi silang 5 kali lipat akan memproses data *training* dan *testing* sebanyak lima kali, sedangkan validasi silang 10 kali lipat akan memproses data *training* dan *testing* sebanyak sepuluh kali Dilakukan proses tersebut untuk mendapatkan nilai akurasi yang tertinggi dalam proses pengujian. Pada proses evaluasi untuk mengukur dengan menggunakan tabel *confusion matrix* meliputi nilai akurasi, nilai *precission*, dan nilai *recall* pada tiap kelas *tweet*.

2.8. Penarikan Kesimpulan

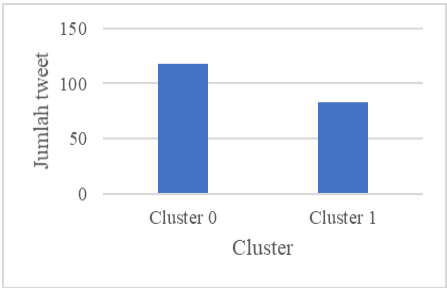
Penarikan kesimpulan berdasarkan hasil kelas *tweet* positif mengandung SARA dan *tweet* negatif mengandung SARA yang didapat pada proses *clustering k-means* serta perhitungan evaluasi dengan menggunakan tabel *confusion matrix* meliputi nilai akurasi, nilai *precission*, dan nilai *recall* yang diperoleh berdasarkan algoritma *Support Vector Machine* (SVM). Gambar 4 menunjukkan seluruh proses pada penelitian ini.

3. HASIL dan PEMBAHASAN

Hasil dari *preprocessing* merupakan kata dasar yang penting dari masing-masing tweet, sedangkan hasil dari proses pembobotan merupakan bobot masing-masing kata pada masing-masing *tweet*. Bobot pada masing-masing kata akan digunakan proses *clustering*, proses *clustering* menggunakan 2 *cluster* yang terdiri dari *cluster* 0 berarti positif SARA dan cluster 1 berarti negatif SARA. Hasil yang diperoleh pada proses ini antara lain *cluster* 0 berjumlah data 118 *tweet* dan *cluster* 1 berjumlah data 83 tweet.



Gambar 4. Metode yang diusulkan



Gambar 5. Hasil proses *clustering* dengan K-means

Data yang telah diolah di *clustering* k-*means* diolah kembali untuk memperbaiki kelas *tweet*. Hasil *clustering* yang sudah didapatkan kemudian divalidasi oleh pakar bahasa, hasil yang diperoleh pada proses ini antara lain *cluster* 0 berjumlah data 139 *tweet* dan *cluster* 1 berjumlah data 62 *tweet*. Sedangkan berdasarkan Gambar 8, hasil proses *clustering* k-means, menunjukan *cluster* 0 memiliki hasil *tweet* terbanyak dibandingkan dengan *cluster* 1 sehingga sentimen *tweet* banyak mengandung unsur SARA. Berikut kalimat-kalimat yang menunjukkan perubahan *cluster* pada *tweet* seperti pada Tabel 2.

Gambar 6. Hasil proses *clustering* yang divalidasi pakar

Untuk menggambarkan secara visualisasi kata apa yang dominan baik pada cluster 0 maupun *cluster* 1, peneliti mengunakan *cloud word*. Visualisasi Cloud Word pada masing-masing *cluster* dapat dilihat pada Gambar 9 dan 10. Terlihat jelas bahwa baik dari *cluster* 0 maupun *cluster* 1, kata yang paling dominan adalah mengenai agama, dan terutama agama yang paling disoroti pada kasus SARA adalah agama islam.

Gambar 7. *Cloud word* positif SARA

Data yang sudah menjadi kelompok, kelompok *tweet* yang mengandung SARA dan *tweet* yang tidak mengandung unsur SARA dilatih untuk menghasilkan pola dari kelas tersebut dengan menggunakan algoritme *Support Vector Machine* (SVM).

Proses klasifikasi algoritme *support vector machine* dengan memaksimalkan batas *hyperplane*. Konsep klasifikasi dengan *support vector machine* untuk mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas data yaitu positif SARA dan negatif SARA.

Proses analisis *tweet* yang telah dilakukan pada proses pre-*processing*, *clustering k-means*, klasifikasi *support vector machine*, tahapan selanjutnya proses validasi dan evaluasi. Pada proses pengujian dilakukan proses validasi *k-fold cross validation*. Proses pengujian menggunakan 5-*fold cross validation* dan 10-*fold cross validation.* Berikut tabel *confusion matrix* pada 5-*fold cross validation*.

Tabel 2. Perubahan Cluster pada *Tweet*

Tweet	*Clustering* k-means	*Clustering* k-means + pakar
Siapa saja yang dukung penista agama adalah bajingan yang perlu diludahi mukanya	Negatif	Positif
Allah kan bukan orang Arab. Tentu Allah senang kalau aya-ayatNya dibaca dg gaya Minang, Ambon, Cina, Hiphop	Negatif	Positif
Ngajari at ngajak org Sunda ngaji, it sama aja, ngajari anjing menggonggong	Negatif	Positif

Gambar 8. *Cloud word* negatif SARA

Tabel 3. *Confusion Matrix 5-Fold Cross Validation*.

	True positif SARA	True negatif SARA
Pred. positif SARA	117	71
Pred. negatif SARA	1	12

Berdasarkan Tabel 3 evaluasi dengan menggunakan *confusion matrix* menghasilkan perhitungan *precision*, *recall*, F-*measure*, dan *accuracy*. Hasil perhitungan yang diperoleh dengan menggunakan *confusion matrix* disajikan dalam Tabel 4 sebagai berikut:

Tabel 4. Performa *Confusion Matrix* 5-Fold *Cross Validation*

Performa	Nilai
Precision	62,23%
Recall	99,15%
F-measure	76,46%
Accuracy	64,18%

Data yang telah diolah di *clustering k-means* 5-*fold cross validation* diolah kembali dengan divalidasi oleh pakar bahasa untuk memperbaiki hasil kelas tiap *tweet.* Hasil yang diperoleh dari proses tersebut disajikan pada Tabel 5 sebagai berikut:

Tabel 5. *Confusion Matrix 5-Fold Cross Validation* dan Validasi Pakar

	True positif SARA	True negatif SARA
Pred. positif SARA	139	60
Pred. negatif SARA	0	2

Berdasarkan tabel 5 evaluasi dengan menggunakan *confusion matrix* menghasilkan *precision*, *recall*, F-*measure*, dan *accuracy*. Hasil perhitungan yang diperoleh dengan menggunakan *confusion matrix* disajikan dalam Tabel 6 sebagai berikut:

Tabel 6. Performa *Confusion Matrix* 5-*fold Cross Validation* Dan Validasi Pakar

Performa	Nilai
Precision	69,85%
Recall	100%
F-measure	82,24%
Accuracy	70,15%

Selanjutnya *k-fold cross validation* yang digunakan adalah 10 . Berikut tabel *confusion matrix* pada 10-*fold cross validation*.

Tabel 7. *Confusion Matrix 10-Fold Cross Validation*

	True positif SARA	True negatif SARA
Pred. positif SARA	117	72
Pred. negatif SARA	1	11

Berdasarkan Tabel 7 evaluasi dengan menggunakan *confusion matrix* menghasilkan perhitungan *precision*, *recall*, *F-measure*, dan *accuracy*. Hasil perhitungan yang diperoleh dengan menggunakan *confusion matrix* disajikan dalam Tabel 8 sebagai berikut:

Tabel 8. Performa *Confusion Matrix 10-Fold Cross Validation*

Performa	Nilai
Precision	61,90%
Recall	99,15%
F-measure	76,21%
Accuracy	63,68%

Data yang telah diolah di *clustering k-means* 10-*fold cross validation* diolah kembali dengan divalidasi oleh pakar bahasa untuk memperbaiki hasil kelas tiap *tweet.* Hasil yang diperoleh dari proses tersebut disajikan pada Tabel 9 sebagai berikut:

Tabel 9. *Confusion Matrix 10-Fold Cross Validation* Dan Manual

	True positif SARA	True negatif SARA
Pred. positif SARA	139	58
Pred. negatif SARA	0	4

Berdasarkan tabel 9 evaluasi dengan menggunakan *confusion matrix* menghasilkan perhitungan *precision*, *recall*, F-*measure*, dan *accuracy*. Hasil perhitungan yang diperoleh dengan menggunakan *confusion matrix* disajikan dalam tabel 10 sebagai berikut:

Tabel 10.. Performa *Confusion Matrix* 10-*Fold Cross Validation* Dan Validasi Pakar

Performa	Nilai
Precision	70,56%
Recall	100%
F-measure	82,74%
Accuracy	71,14%

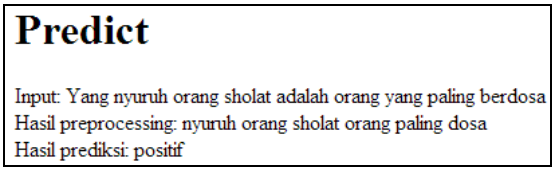
Berdasarkan hasil analisis yang telah dilakukan maka untuk mengevaluasi model yang diperoleh dari hasil ekperimen didapatkan sebagai berikut:

Tabel 1. Performa *Confusion Matrix* 10-*Fold Cross Validation* Dan Validasi Pakar

K-Fold Cross Validation	Model	*Precision*	*Recall*	*F-measure*	Akurasi
5-Fold Cross Validation	***K-means* + SVM**	62,23%	99,15%	76,18%	64,18%
	***K-means* + validasi pakar + SVM**	69,85%	100%	82,24%	70,15%
10-Fold Cross Validation	***K-means* + SVM**	61,90%	99,15%	76,21%	63,68%
	***K-means* + validasi pakar + SVM**	70,56%	100%	82,74%	71,14%

Dari hasil Tabel 10 diperlihatkan bahwa setelah dilakukan eksperimen terdapat perbedaan antara model dari *k-means* + *SVM* dibandingkan dengan model *k-means* + validasi pakar + SVM mengalami peningkatan. Tabel 10 menunjukkan hasil nilai *precision* tertinggi sebesar 70,56% pada model *k-means* + validasi pakar + *SVM* 10-*fold cross validation*, nilai *recall* tertinggi sebesar 100% pada model *k-means* + validasi pakar + *SVM* 5-*fold cross validation* dan *k-means* + validasi pakar + *SVM* 10-*fold cross validation*, nilai *f-measure* tertinggi sebesar 82,74% pada model *k-means* + validasi pakar + *SVM* 10-*fold cross validation*, dan nilai akurasi tertinggi sebesar 71,14% pada model *k-means* + validasi pakar + *SVM* 10-*fold cross validation*.

Simulasi atau *prototype* digunakan untuk mendukung hasil *preprocessing* dan hasil prediksi kelas dengan memasukan *tweet* yang ingin diseleksi. Seperti pada Gambar 9.



Gambar 9. Simulasi prediksi

4. KESIMPULAN

Berdasarkan hasil dari penelitian yang telah dilakukan, Kombinasi *k-means* dan *support vector machine* (SVM) sudah berhasil untuk menganalisis *tweet* yang positif mengandung SARA dan *tweet* yang negatif mengandung SARA akan tetapi Hasil pengelompokan *tweet* yang mengandung SARA dan tidak mengandung SARA belum memiliki hasil yang baik, dibuktikan dengan hasil klasifikasi menggunakan algoritme SVM masih rendah. Setelah proses pengelompokan diperbaiki dengan validasi pakar bahasa, hasil akurasi dari proses klasifikasi dapat meningkat baik dengan 5-*fold cross validation* maupun 10-*fold cross validation*. Sehingga untuk penelitian selanjutnya perlu adanya perbaikan metode yang disulkan untuk memperoleh hasil yang lebih akurat, baik dalam mengelompokan dan klasifikasi data twitter yang mengandung unsur SARA dan tidak mengandung unsur SARA. Selain itu, sebaiknya dibangun sistem yang dapat diterapkan untuk menganalisis konten twitter.

DAFTAR PUSTAKA

DUWAIRI, R. dan ABU-RAHMEH, M., 2015. A novel approach for initializing the spherical K-means clustering algorithm. *Simulation Modelling Practice and Theory*, [online] 54, pp.49–63. Available at: <http://dx.doi.org/10.1016/j.simpat.2015.03.007>.

GAO, J., CHENG, Q. & YU, P.L.H., 2019. Detecting Comments Showing Risk for Suicide in YouTube. In: *Proceedings of the Future Technologies Conference*. [online] Springer, Cham, pp.385–400. Available at: <http://link.springer.com/10.1007/978-3-030-02683-7>.

GARAY, J., YAP, R. & SABELLANO, M.J., 2019. An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer. *IOP Conference Series: Materials Science and Engineering*, 482, pp.1–6.

GULNASHIN, F., SHARMA, I. & SHARMA, H., 2019. Progress in Advanced Computing and Intelligent Engineering. *Progress in Advanced Computing and Intelligent Engineering*, [online] 714, pp.149–155. Available at: <http://link.springer.com/10.1007/978-981-13-0224-4>.

HAN, J., KAMBER, M. & PEI, J., 2012. *Data mining: concepts* and *techniques*. Third Edit ed. [online] *Vasa*. London, UK: Morgan Kaufmann. Available at: <http://medcontent.metapress.com/index/A65RM03P4874243N.pdf>.

HOOTSUITE dan WE ARE SOCIAL, 2018. *Essential Insights Into Internet, Social Media, Mobile, and E*-Commerce *Use Around The World*. [online] Digital in 2018. Available at: <https://wearesocial.com/blog/2018/01/global-digital-report-2018> [Accessed 25 Jun. 2018].

JOACHIMS, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *European Conference on Machine Learning*. [online] pp.137–142. Available at: <http://www.springerlink.com/index/drhq581108850171.pdf>.

KE, B., SHEN, X.D., JI, H., GAO, F., KAMO, N., ZHAI, Y., BUSUTTIL, R. V & KUPIEC-WEGLINSKI, J.W., 2011. STAT3-PTEN Axis: A Negative Regulator of Dendric Cell-Mediated Innate Immune Functions. *American Journal of Transplantation*, 11, p.197.

KEMENTERIAN KOMUNIKASI DAN INFORMATIKA, 2017. *Ujaran Kebencian Picu Generasi Muda Jadi Intoleran dan Diskriminatif*. Jakarta.

KOROVKINAS, K., DANĖNAS, P. & GARŠVA, G., 2019. SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis. *Baltic Journal of Modern Computing*, 7(1).

LIU, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, [online] 5(1), pp.1–167. Available at: <http://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016>.

LIU, S. dan LEE, I., 2018. Email Sentiment Analysis Through k-Means Labeling and Support Vector Machine Classification. *Cybernetics and Systems*, 49(3), pp.181–199.

MA, S. dan CHEN, X., 2019. A data mining approach to predict risk of cardiovascular A Data Mining Approach to Predict Risk of Cardiovascular. In: *AIP Conference Proceedings*. pp.020014-1-020014–7.

MANEVITZ, L.M. dan YOUSEF, M., 2001. One-Class SVMs for Document Classification.

Journal of Machine Learning Research 2, 2, pp.139–154.

MEDHAT, W., HASSAN, A. & KORASHY, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, [online] 5(4), pp.1093–1113. Available at: <http://dx.doi.org/10.1016/j.asej.2014.04.011>.

MUSTAKIM, 2012. Pemetaan Digital dan Pengelompokan Lahan Hijau di Wilayah Provinsi Riau Berdasarkan Knoledge Discovery in Databases (KDD) dengan Teknik K-Means Mining. In: *Seminar nasional Teknologi Informasi Komunikasi dan Industri (SNTIKI*. Pekanbaru, Riau: Fakultas Sains dan Teknologi UIN Sultan Syarif Kasim Riau, pp.103–111.

MUZAKIR, A., 2014. Analisa Dan Pemanfaatan Algoritma K-Means Clustering pada Data Nilai Siswa sebagai Penentuan Penerimaan Beasiswa. In: *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*. Yogyakarta: Institut Sains & Teknologi Yogyakarta, p.A-195-A-200.

PRASETYO, E., 2014. *Data Mining, Mengolah Data Menjadi* Informasi *Menggunakan Matlab*. 1st Publis ed. Yogyakarta: Andi Offset.

RAHMAWATI, A., MARJUNI, A. & ZENIARJA, J., 2017. Analisis Sentimen Publik Pada Media Sosial Twitter Terhadap Pelaksanaan Pilkada Serentak Menggunakan Algoritma Support Vector Machine. *CCIT Journal*, [online] 10(2), pp.197–206. Available at: <http://ejournal.raharja.ac.id/index.php/ccit/article/view/67>.

ROZI, N.F., ARIANTO, F. & HAPSARI, D.P., 2019. Analisis Sentimen Pada Opini Pengguna Maskapai Penerbangan Sentiment Analysis on Passenger Opinions At Airlines Company. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, [online] 6(3), pp.321–326. Available at: <http://jtiik.ub.ac.id/index.php/jtiik/article/view/1337/pdf>.

RUDYBYO, 2011. *Pengertian Sara: Suku, Ras, Agama, dan Antaragama*. [online] Available at: <http://rudybyo.blogspot.co.id/2011/04/pengertian-sara-suku-ras-agama-dan.html> [Accessed 10 Oct. 2018].

SOMANTRI, O. dan APRILIANI, D., 2018. Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Support Vector Machine Based on Feature Selection for Sentiment Analysis Customer Satisfaction on Culinary. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, [online] 5(5), pp.537–548. Available at: <http://jtiik.ub.ac.id/index.php/jtiik/article/view/867/pdf>.

SOMANTRI, O., WIYONO, S. & DAIROH, , 2016. Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM). *Scientific Journal of Informatics Universitas Negeri Semarang*, 3, pp.34–45.

TIM PUSAT HUMAS KEMENTERIAN PERDAGANGAN RI, 2014. *Panduan Optimalisasi Media Sosial untuk Kementerian Perdagangan RI*. [online] Tim Pusat Humas Kementerian Perdagangan RI. Available at: <http://www.kemendag.go.id/files/pdf/2015/01/15/buku-media-sosial-kementerian-ido-1421300830.pdf> [Accessed 21 May 2018].

XU, D. dan TIAN, Y., 2015. A Comprehensive Survey of *Clustering* Algorithms. *Annals of Data Science*, 2(2), pp.165–193.

ZHANG, W., YOSHIDA, T. & TANG, X., 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, [*online*] 38(3), pp.2758–2765. Available at: <http://dx.doi.org/10.1016/j.eswa.2010.08.066>.

Halaman ini sengaja dikosongkan